IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, November 2025
Impact Factor: 7.67

AI-Driven Virtual Assistant Transforming Image into Voice

Chaithra KR[1] and Prof. Thouseef Ulla Khan[2]
Student, Department of MCA[1]
Assistant Professor, Department of MCA[2]
Vidya Vikas Institute of Engineering and Technology, Mysuru

Abstract: *This project focuses on developing A system capable of automatically producing captions for pictures and convert those captions into natural-sounding speech. The system leverages The Flickr8K dataset, which is contains thousands of photos with numerous captions added by humans,to train a model for deep learning. The visual characteristics of the images are extracted using a pre-trained VGG16 Neural Network Convolution (CNN), while the sequence generation is performed using Networks using Long Short-Term Memory (LSTM) combined with a system of attention to produce contextually accurate and grammatically correct captions.The generated captions are then passed to a Text-to-Speech engine, which converts the passage into audio. The incorporation of vision, language, and speech technologies results within a structure that is not just able to describe images but also narrates them aloud, making it very helpful for situations such as assisting visually impaired individuals, automated content creation, education tools, and interactive systems.The project is implemented using TensorFlow/Keras for deep learning and Flask for building a web-based front end. The interface allows users to upload an image, view the generated caption, and listen to the audio narration. The evaluation of The apparatus is done using BLEU scores, which measure the caliber of the produced captions against human-annotated references. While the system achieves reasonable performance, challenges remain in improving accuracy, handling complex images, and producing more natural voice outputs.*

Keywords*:* Image Captioning, Image-to-Voice System, Convolutional Neural Network (CNN), VGG16, Long Short-Term Memory (LSTM), Attention Mechanism, Text-to-Speech (TTS), Flickr8k Dataset, Assistive Technology, Multimodal AI, Flask Web Application

DOI: 10.48175/IJARSCT-30026